WHAT IS CLAIMED IS:

Sub A1

1. A document segmentation apparatus comprising:

table analyzing means for generating cell position data indicating a positional relationship between cells and cell vectors representing characteristics of the cells, by analyzing a table in a document to be processed;

table type judging means for judging a table type with reference to the cell position data and the cell vectors generated by said table analyzing means;

first segment generating means for generating a segment from the table when the table type is a table for showing a table; and

second segment generating means for generating a segment from the table when the table type is a table for layout.

2. A document segmentation apparatus according to claim 1, wherein said first segment generating means comprise;

cut direction determination means for determining a cut direction of the table by judging whether the data is expressed in a column or a row in the table on the basis of the cell position data and the cell vectors; and

table segment generating means for generating a table segment by dividing the table on the basis of the table type and the cut direction.

3. A document segmentation apparatus according to claim 2, wherein said second segment generating means generate the table itself as the segment.

4. A document segmentation apparatus according to claim 1, wherein said second segment generating means comprise;

cell cluster generating means for generating cell cluster information by clustering the cells in the table; and

layout segment generating means for generating segment by connecting the cells in the table with reference to the cell position data and the cell cluster information.

5. A document segmentation apparatus according to claim 4, wherein said first segment generating means generate the table itself as the segment.

6. A document segmentation apparatus according to claim 4, wherein said second segment generating means generate the table itself as the segment.

7. A document segmentation apparatus according to claim 1, further comprising normal segment

generating means for dividing the document into a segment which corresponds to one table;
and wherein

the table generated as one segment by said normal segment generating means is to be processed by said table analyzing means.

8. A document segmentation apparatus according to claim 1, wherein said table analyzing means further generate cell data of the analyzed table and said table type judging means judge the table type with reference to the cell data.

9. A document segmentation apparatus according to claim 8, wherein said table type judging means comprise similarity judging means for judging the table type on the basis of similarity between the cell data positioned at particular positions with reference to the cell position data and the cell data generated by said table analyzing means.

10. A document segmentation apparatus according to claim 8, wherein said table type judging means comprise partial character line extracting means for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by

said table analyzing means, and character line comparing means for comparing the extracted partial character lines to judge the table type.

11. A document segmentation apparatus according to claim 8, wherein said table type judging means comprise partial character line extracting means for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by said table analyzing means, and similarity judging means for judging the table type on the basis of similarity between the extracted partial character lines.

12. A document segmentation apparatus according to claim 8, wherein said table type judging means comprise syntax judging means for judging the table type with reference to the cell position data, the cell vectors and the cell data generated by said table analyzing means, and similarity judging means for judging the table type on the basis of similarity between the cell data positioned at particular positions with reference to the cell position data and the cell data generated by said table analyzing means.

13. A document segmentation apparatus according

to claim 8, wherein said table type judging means comprise syntax judging means for judging the table type with reference to the cell position data, the cell vectors and the cell data generated by said table analyzing means, partial character line extracting means for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by said table analyzing means, and character line comparing means for comparing the extracted partial character lines to judge the table type.

14. A document segmentation apparatus according to claim 8, wherein said table type judging means comprise syntax judging means for judging the table type with reference to the cell position data, the cell vectors and the cell data generated by said table analyzing means, partial character line extracting means for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by said table analyzing means, and similarity judging means for judging the table type on the basis of similarity between the extracted partial character lines.

15. A document segmentation apparatus according

to claim 1, further comprising table reforming means for reforming the table so that the number of cells in each column and each row becomes the same, by analyzing the table to be processed;

and wherein

said table analyzing means analyze the reformed table.

16. A document segmentation apparatus according to claim 15, wherein said table reforming means comprise supplementary data removing means for removing data added to the table from the table data.

17. A document segmentation apparatus according to claim 15, wherein said table reforming means comprise multi-row/multi-column processing means for reforming the table regularly by analyzing the structure of the table data.

18. A document segmentation apparatus according to claim 15, wherein said table reforming means comprise composite table processing means for reforming the table by analyzing regularity of information description constituting the table.

19. A document segmentation apparatus according to claim 15, wherein said table reforming means

comprise;

supplementary data removing means for removing data added to the table from the table data; and

multi-row/multi-column processing means for reforming the table regularly by analyzing the structure of the table data.

20. A document segmentation apparatus according to claim 15, wherein said table reforming means comprise;

supplementary data removing means for removing data added to the table from the table data; and

composite table processing means for reforming the table by analyzing regularity of information description constituting the table.

21. A document segmentation apparatus according to claim 15, wherein said table reforming means comprise;

multi-row/multi-column processing means for reforming the table regularly by analyzing the structure of the table data; and

composite table processing means for reforming the table by analyzing regularity of information description constituting the table.

22. A document segmentation apparatus according to claim 15, wherein said table reforming means comprise:

supplementary data removing means for removing data added to the table from the table data;

multi-row/multi-column processing means for reforming the table regularly by analyzing the structure of the table data; and

composite table processing means for reforming the table by analyzing regularity of information description constituting the table.

23. A document segmentation method comprising:

a table analyzing step for generating cell position data indicating a positional relationship between cells and cell vectors representing characteristics of the cells, by analyzing a table in a document to be processed;

a table type judging step for judging a table type with reference to the cell position data and the cell vectors generated by said table analyzing step;

a first segment generating step for generating a segment from the table when the table type is a table describing a table; and

a second segment generating step for generating a segment from the table when the table type is a table for layout.

24. A document segmentation method according to claim 23, wherein said first segment generaitng step comprises:

a cut direction determination step for determining a cut direction of the table by judging whether the data is expressed in a column or a row in the table on the basis of the cell position data and the cell vectors; and

a table segment generating step for generating a table segment by dividing the table on the basis of the table type and the cut direction.

25. A document segmentation method according to claim 24, wherein said second segment generating step generates the table itself as the segment.

26. A document segmentation method according to claim 23, wherein said second segment generating step comprises;

a cell cluster generating step for generating cell cluster information by clustering the cells in the table; and

a layout segment generating step for generating segment by connecting the cells in the table with reference to the cell position data and the cell cluster information.

27. A document segmentation method according to claim 26, wherein said first segment generating step generates the table itself as the segment.

28. A document segmentation method according to claim 26, wherein said second segment generating step generates the table itself as the segment.

29. A document segmentation method according to claim 23, further comprising a normal segment generating step for dividing the document into a segment which corresponds to one table;
and wherein
the table generated as one segment by said normal segment generating step is to be processed by said table analyzing step.

30. A document segmentation method according to claim 23, wherein said table analyzing step further generates cell data of the analyzed table and said table type judging step judges the table type with reference to the cell data.

31. A document segmentation method according to claim 30, wherein said table type judging step

comprises a similarity judging step for judging the table type on the basis of similarity between the cell data positioned at particular positions with reference to the cell position data and the cell data generated by said table analyzing step.

32. A document segmentation method according to claim 30, wherein said table type judging step comprises a partial character line extracting step for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by said table analyzing step, and a character line comparing step for comparing the extracted partial character lines to judge the table type.

33. A document segmentation method according to claim 30, wherein said table type judging step comprises a partial character line extracting means for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by said table analyzing step, and a similarity judging step for judging the table type on the basis of similarity between the extracted partial character lines.

34. A document segmentation method according to claim 30, wherein said table type judging step comprises a syntax judging step for judging the table type with reference to the cell position data, the cell vectors and the cell data generated by said table analyzing step, and a similarity judging step for judging the table type on the basis of similarity between the cell data positioned at particular positions with reference to the cell position data and the cell data generated by said table analyzing step.

35. A document segmentation method according to claim 30, wherein said table type judging step comprises a syntax judging step for judging the table type with reference to the cell position data, the cell vectors and the cell data generated by said table analyzing step, a partial character line extracting step for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by said table analyzing step, and a character line comparing step for comparing the extracted partial character lines to judge the table type.

36. A document segmentation method according to claim 30, wherein said table type judging step

comprises a syntax judging step for judging the table type with reference to the cell position data, the cell vectors and the cell data generated by said table analyzing step, a partial character line extracting step for extracting partial character lines from the cell data positioned at a particular position with reference to the cell position data and the cell data generated by said table analyzing step, and a similarity judging means for judging the table type on the basis of similarity between the extracted partial character lines.

37. A document segmentation method according to claim 23, further comprising a table reforming step for reforming the table so that the number of cells in each column and each row becomes the same, by analyzing the table to be processed;
and wherein
said table analyzing step analyzes the reformed table.

38. A document segmentation method according to claim 37, wherein said table reforming step comprises a supplementary data removing step for removing data added to the table from the table data.

39. A document segmentation method according to

claim 37, wherein said table reforming step comprises a multi-row/multi-column processing step for reforming the table regularly by analyzing the structure of the table data.

40. A document segmentation method according to claim 37, wherein said table reforming step comprises a composite table processing step for reforming the table by analyzing regularity of information description constituting the table.

41. A document segmentation method according to claim 37, wherein said table reforming step comprises;

a supplementary data removing step for removing data added to the table from the table data; and

a multi-row/multi-column processing step for reforming the table regularly by analyzing the structure of the table data.

42. A document segmentation method according to claim 37, wherein said table reforming step comprises;

a supplementary data removing step for removing data added to the table from the table data; and

a composite table processing step for reforming the table by analyzing regularity of information description constituting the table.

43. A document segmentation method according to claim 37, wherein said table reforming step comprises;

a multi-row/multi-column processing step for reforming the table regularly by analyzing the structure of the table data; and

a composite table processing step for reforming the table by analyzing regularity of information description constituting the table.

44. A document segmentation method according to claim 37, wherein said table reforming step comprises;

a supplementary data removing step for removing data added to the table from the table data;

a multi-row/multi-column processing step for reforming the table regularly by analyzing the structure of the table data; and

a composite table processing step for reforming the table by analyzing regularity of information description constituting the table.

45. A computer-readable storage medium storing a document segmentation program for controlling a computer to perform document segmentation, said program comprising codes for causing the computer to perform:

a table analyzing step for generating cell position data indicating a positional relationship between cells and cell vectors representing characteristics of the cells, by analyzing a table in a document to be processed;

a table type judging step for judging a table type with reference to the cell position data and the cell vectors generated by said table analyzing step;

a first segment generating step for generating a segment from the table when the table type is a table describing a table; and

a second segment generating step for generating a segment from the table when the table type is a table for layout.